e-ISSN: 2987-1530

Sriwijaya Journal of Forensic and Medicolegal (SJFM)

Journal website: https://phlox.or.id/index.php/sjfm

Neural Network Versus Stepwise Regression for Forensic Stature Estimation from Percutaneous Tibial Dimensions in South Sumatran Malay Adults

Sari Sulistyoningsih[1], Abu Bakar[2], Eduardo Michael Perez[3], Ifah Shandy[4]*

[1]Department of Anatomy, CMHC Research Center, Palembang, Indonesia

[2]Department of Radiology, CMHC Research Center, Palembang, Indonesia

[3]Department of Oral Biology, Cobija State Hospital, Cobija, Bolivia

[4]Department of Mathematics Education, Enigma Institute, Palembang, Indonesia

ARTICLE INFO

Keywords:
Artificial neural network
Forensic anthropology
Malay population
Stature estimation
Tibia

***Corresponding author:**
Ifah Shandy

E-mail address:
ifah.shandy@enigma.ac.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.59345/sjfm.v3i2.248

A B S T R A C T

Introduction: Stature estimation from skeletal elements underpins forensic biological profiling and is decisive in disaster victim identification, yet no validated osteometric standard exists for the South Sumatran Malay population, and stepwise regression assumes a linearity that skeletal growth biology does not obey.

Methods: 450 healthy South Sumatran Malay adults (225 male, 225 female) aged 20-50 years were enrolled at CMHC Research Center, Palembang, and reported per STROBE. Five percutaneous right-tibial dimensions were measured by one certified anthropologist under the Martin and Saller convention (all relative technical errors of measurement below 1.5%), and a stratified 70:30 partition withheld 135 observations for testing. Stepwise multiple linear regression (SMLR) was compared against a grid-search-optimised 5-64-32-16-1 multilayer perceptron artificial neural network (MLP-ANN).

Results: Sexual dimorphism was large for every dimension (Cohen's d 1.42-1.75; all $p < 0.001$). Percutaneous tibial length was the strongest single predictor (males $r = 0.812$, 95% CI 0.762-0.852). The calibration-corrected pooled regression reached $R^2 = 0.742$ with RMSE ±4.82 cm, whereas the network reached $R^2 = 0.914$ (95% CI 0.891-0.933) and RMSE ±2.78 cm — $\Delta R^2 = 0.172$ (bootstrap 95% CI 0.141-0.203), a 23.2% gain in explained variance and a 42.3% reduction in error that narrows the 95% forensic identification window from ±9.45 cm to ±5.45 cm.

Conclusion: These first population-specific, artificial-intelligence-driven standards materially improve the biological profiling of incomplete human remains in Indonesian medicolegal and disaster victim identification practice.

1. Introduction

The reconstruction of living stature from skeletal elements is one of the most consequential analytical tasks in forensic anthropological practice. Within the operational framework of disaster victim identification (DVI), an accurate stature estimate derived from fragmentary or skeletonised remains provides a quantitative biometric anchor for matching post-mortem findings against ante-mortem records, including national identity documentation, military enlistment files, occupational health registers and clinical height measurements.[1] Indonesia occupies a geographic setting characterised by exceptional natural-hazard exposure — volcanic eruption, tsunami, seismic events and aviation accidents recur across the archipelago — and the country's DVI apparatus is correspondingly mature but chronically constrained by the absence of locally derived biological-profile reference data.[2] In this setting, the availability of reliable population-specific standards is not merely a scientific aspiration; it is a medicolegal operational necessity, and one that is explicitly engaged by Articles 133 and 179 of the Indonesian Code of Criminal Procedure (Kitab Undang-Undang Hukum Acara Pidana, KUHAP), which oblige the treating or examining physician to furnish the investigating authority with a formal expert medical opinion.

The mathematical method of stature reconstruction — deriving predictive regression equations from measured long-bone lengths and breadths — has dominated forensic anthropological practice since the foundational work of Trotter and Gleser. Among the long bones available for this purpose the tibia is a particularly valuable element. Its direct contribution to the inferior-limb component of total stature, its dense cortical architecture (which confers superior survival in taphonomically hostile environments), and the practicality of percutaneous measurement in fleshed or partially decomposed remains combine to make it a preferred element for stature reconstruction in both field and laboratory forensic settings.[3,4] Percutaneous

tibial measurement, which captures the external dimensions of the bone through overlying soft tissue using spreading calipers and an osteometric board, carries the additional advantage of being applicable both to intact living subjects for the construction of reference databases and to recently deceased individuals in the immediate post-disaster phase.[5,6]

A fundamental biological principle constrains the universal applicability of any regression equation derived from a single reference population: the allometric relationship between tibial morphometry and total stature is not invariant across ethnic groups. That relationship is jointly determined by genetic factors — including the expression of homeobox developmental genes and the sensitivity of growth-plate chondrocytes to somatotropic signalling — and by environmental determinants including nutritional adequacy during critical growth windows and the mechanical loading imposed by habitual physical activity.[7,8] Populations inhabiting equatorial regions, including the Malay ethnic groups of insular Southeast Asia, exhibit characteristic limb-to-trunk proportions consistent with the thermoregulatory predictions of Bergmann's and Allen's ecological rules, producing limb morphologies that differ systematically from European or Northeast Asian reference groups.[9] Applying non-population-specific equations to a South Sumatran Malay decedent therefore introduces a predictable systematic error that is incompatible with the precision demanded of forensic biological profiling.[10]

The South Sumatran Malay population, residing principally within the drainage basins of the Musi, Ogan, Komering and Lematang rivers and concentrated in the metropolitan area of Palembang, has remained essentially absent from the forensic osteometric literature. This demographic constitutes one of the major ethnic constituencies of the Sumatran mainland and is characterised by an ancestral composition reflecting Austronesian maritime migration, Dravidian and Arab trade contact, and an indigenous Sundaic substrate. The forensic anthropological community has repeatedly identified the development of population-specific standards for this and related Indonesian ethnic groups as a high-priority objective, yet empirically validated equations remain unavailable.[2,11] CMHC Research Center in Palembang, which serves as a referral facility for anatomical, radiological and medicolegal work across South Sumatra Province, is well positioned to generate a reference dataset that is genuinely representative of this population.

Concurrently, forensic anthropology is undergoing a methodological transformation driven by the integration of artificial intelligence and machine learning.[12-14] Stepwise multiple linear regression (SMLR), despite decades of dominance as the primary tool for osteometric equation derivation, operates under a strict assumption of linearity between each predictor and the outcome — an assumption that misrepresents the biological complexity of skeletal growth and morphogenesis. Feed-forward multilayer perceptron artificial neural networks (MLP-ANN) function as universal non-linear function approximators, learning arbitrary mappings between input measurement vectors and continuous outcomes through iterative gradient-based optimisation, without imposing an a priori structural constraint on the functional form of the predictor-outcome relationship.[15] Recent investigations in adjacent biological-profile domains have demonstrated substantial performance advantages for network-based approaches over traditional linear methods. The most directly relevant demonstration is Indonesian: on 200 cranial computed tomography scans, a deep-learning model classified sex with 97% accuracy against 82% for an experienced human assessor,[16] and comparable gains have been reported for virtual pelvic models in a Han Chinese sample.[17] Whether that advantage extends to the continuous-outcome domain of stature estimation, and by how much, has not been quantified in any Indonesian population.

The present study was therefore designed to address a dual scientific gap. Its objectives were (1) to derive the first empirically validated, population-specific stature-estimation equations for the South Sumatran Malay population from five percutaneous tibial dimensions; (2) to train and evaluate an optimised MLP-ANN on the same measurement set and quantify the magnitude of any predictive improvement over SMLR on a withheld test partition; and (3) to characterise the sexual dimorphism of the tibial dimensions concerned, both as a descriptive reference for Indonesian forensic practice and as a secondary indicator of their value for sex estimation. The findings are intended to contribute population-specific reference data to the Indonesian Association of Forensic Medicine Specialists (Perhimpunan Dokter Forensik Indonesia, PDFI), to the forensic units of the Indonesian National Police (POLRI), and to INTERPOL-aligned DVI workflows deployed in the region.[18]

2. Methods

2.1. Study design and reporting standard

A cross-sectional observational anthropometric study was conducted in South Sumatra Province, Indonesia, between January 2024 and December 2024. The study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement for cross-sectional studies,[19] and the model-development and model-evaluation components are additionally reported against the principles of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) framework. The overall study architecture, from enrolment through blinded evaluation on the withheld test partition, is summarised in Figure 1.

2.2. Ethical approval

Ethical clearance for this study was granted by the Institutional Ethics Committee of CMHC Research Center, Palembang, Indonesia (Approval No. 2024/047). Written informed consent was obtained from every participant prior to enrolment. The study was conducted in full conformity with the principles of the Declaration of Helsinki (2013 revision) and with all applicable Indonesian national regulations governing research involving human subjects. Participation was voluntary, no personally identifying information was retained in the analytical dataset, and participants were free to withdraw at any stage without consequence.

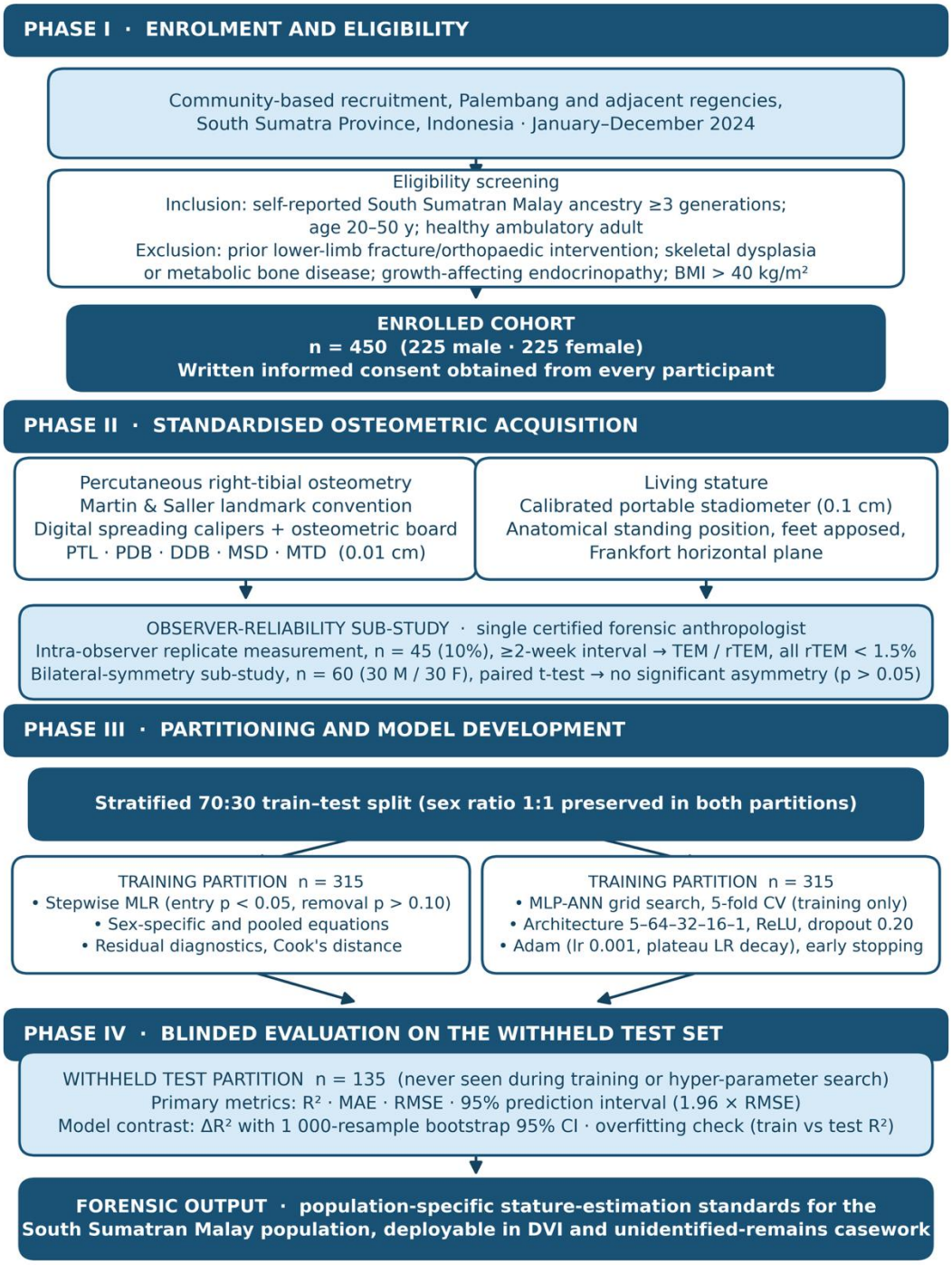

Figure 1. Study design and analytical workflow, from community-based enrolment and eligibility screening through standardised osteometric acquisition, observer-reliability and bilateral- symmetry sub-studies, stratified partitioning and parallel model development, to blinded evaluation on the withheld test partition.

2.3. Setting and participants

The study enrolled 450 healthy adult volunteers (225 male, 225 female) aged 20-50 years, recruited from multiple community settings within Palembang and its adjacent regencies in order to improve representativeness across the occupational and socioeconomic spectrum. Ethnic identity was operationalised as self-reported South Sumatran Malay ancestry for at least three consecutive generations, verified through a structured demographic interview administered by trained research staff.

Mean age was 33.7 ± 8.4 years in males (range 20-50 years) and 34.2 ± 8.1 years in females (range 20-50 years); mean body mass index was 23.4 ± 3.2 kg/m² in males and 22.8 ± 3.1 kg/m² in females. Exclusion criteria comprised prior lower-limb fracture or orthopaedic intervention, congenital skeletal dysplasia or metabolic bone disease, endocrine disorders affecting growth or bone metabolism, and body mass index above 40 kg/m². The eligibility cascade is shown in Figure 1.

2.4. Sample size and statistical power

The target sample size was fixed a priori using a linear-regression power framework implemented in G*Power v.3.1. Assuming an anticipated coefficient of determination of approximately 0.65 for tibial-length-

based stature regression in comparable Southeast Asian populations,[7,20] a minimum acceptable R^2 of 0.50, five predictors, α = 0.05 (two-tailed) and power = 0.90, a minimum of 187 subjects per sex was indicated. The enrolled sample of 225 subjects per sex (total n = 450) exceeds that minimum by 20.3%. Recomputed post hoc for the enrolled sample, the achieved power to detect an R^2 of 0.50 with five predictors exceeds 0.999, and the design is therefore not power-limited for either the SMLR derivation or the SMLR-versus-ANN comparison.

2.5. Osteometric measurement protocol

All measurements were performed by a single certified forensic anthropologist with documented expertise in percutaneous osteometry, using the landmark conventions of Martin and Saller.[21] Living stature was recorded with a calibrated portable stadiometer (precision 0.1 cm) with the subject in the anatomical standing position, feet apposed, and the Frankfort horizontal plane oriented horizontally. Five percutaneous dimensions of the right tibia were then acquired with calibrated digital spreading calipers and an osteometric board, each recorded to the nearest 0.01 cm. The measurement definitions and their landmark-to-landmark derivation are illustrated in Figure 2.

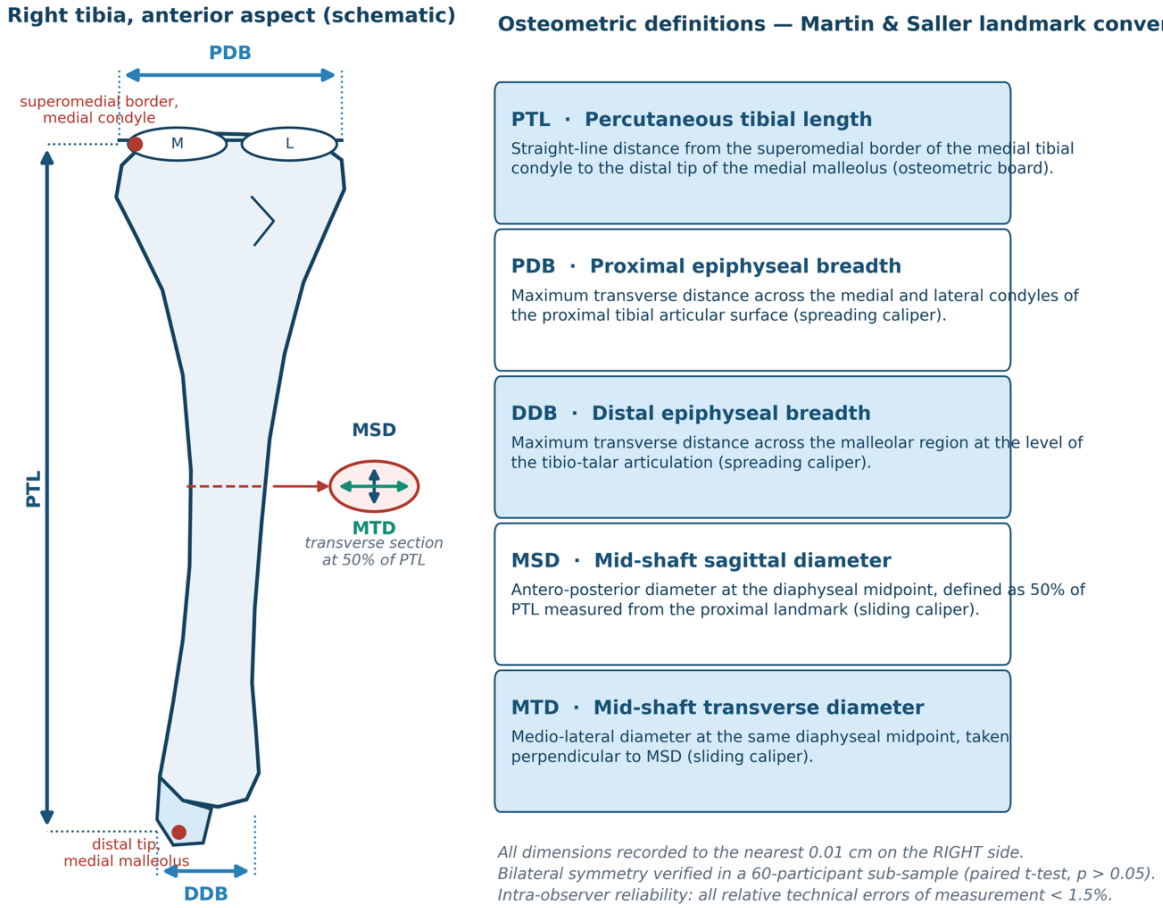

Figure 2. Percutaneous tibial osteometry. Schematic of the right tibia in anterior aspect showing the landmark-to-landmark derivation of the five measured dimensions under the Martin and Saller convention, with the transverse section at 50% of percutaneous tibial length from which the mid-shaft diameters were taken, and the operational definition of each dimension.

Percutaneous tibial length (PTL) is the straight-line distance from the superomedial border of the medial tibial condyle to the distal tip of the medial malleolus. Proximal epiphyseal breadth (PDB) is the maximum transverse distance across the medial and lateral condyles of the proximal tibial articular surface. Distal epiphyseal breadth (DDB) is the maximum transverse distance across the malleolar region at the level of the tibio-talar articulation. Mid-shaft sagittal diameter (MSD) is the antero-posterior diameter at the diaphyseal midpoint, defined as 50% of PTL measured from the proximal landmark, and mid-shaft transverse diameter (MTD) is the medio-lateral diameter at the same point, taken perpendicular to MSD. Each landmark pair is annotated on the schematic in Figure 2.

2.6. Measurement reliability and laterality

Intra-observer reliability was established by repeat measurement of 45 randomly selected participants (10% of the cohort) at a minimum interval of two weeks. The technical error of measurement (TEM) and the relative technical error of measurement (rTEM) were computed for every variable. All rTEM values fell below 1.5%, comfortably within the 5% threshold conventionally accepted for anthropometric reliability,[22,23] confirming that measurement error is small relative to the between-subject variance the models are required to explain. Per-variable TEM values are held in the study's primary records and are available from the corresponding author on reasonable request.

Bilateral symmetry was assessed in a sub-sample of 60 participants (30 male, 30 female) in whom both tibiae were measured. No statistically significant left-right asymmetry was identified for any variable (paired t-test, $p > 0.05$ throughout), which supports the use of right-sided measurements as the study standard and, by extension, the applicability of the derived equations to either tibia in casework.

2.7. Model development and partitioning

A stratified 70:30 train-test split was applied, allocating 315 observations to model training and 135 observations to model evaluation, with stratification preserving the 1:1 sex ratio within each partition. The test partition was withheld from every stage of model specification, variable selection and hyper-parameter search.

For SMLR, all five tibial dimensions were entered into a stepwise selection procedure with an entry criterion of $p < 0.05$ and a removal criterion of $p > 0.10$. Sex-specific and pooled models were derived. Regression coefficient 95% confidence intervals and individual predictor t-statistics were computed. Residual diagnostics comprised plots of residuals against fitted values and normal quantile-quantile plots of standardised residuals, assessing linearity, homoscedasticity and residual normality; Cook's distance was computed to identify influential observations.

For the MLP-ANN, hyper-parameter optimisation was conducted exclusively within the training partition using five-fold cross-validation and grid search, so that no information from the test partition could leak into model selection. The optimised architecture comprised an input layer of five neurons, three hidden layers of 64, 32 and 16 neurons with rectified linear unit activation, and a single continuous output neuron. The Adam optimiser was used with an initial learning rate of 0.001 and a schedule halving the learning rate on validation-loss plateau (patience 20 epochs). Dropout regularisation (rate 0.20) was applied to each hidden layer during training, the batch size was 32, and training was terminated by early stopping on minimum validation loss (patience 50 epochs) with a ceiling of 500 epochs.

2.8. Statistical analysis

Analyses were performed in SPSS v.26.0 (IBM Corp., Armonk, NY, USA) and in Python v.3.11 using the scikit-learn, TensorFlow, SciPy and NumPy libraries. Distributional normality was assessed by the Shapiro-Wilk test. Descriptive statistics are presented as mean ± standard deviation with the observed range.

Sexual dimorphism was evaluated by independent-samples t-test under a pooled-variance model ($df = 448$); exact t-statistics and p-values are reported to three decimal places, and the familywise error rate across the six simultaneous comparisons was controlled by Bonferroni correction (adjusted threshold $p < 0.008$). Effect size is reported as Cohen's d with a 95% confidence interval, with Hedges' g as the small-sample-corrected analogue and the variance-explained analogue η^2 for each contrast. A sexual dimorphism index, expressed as the male mean as a percentage of the female mean, is reported for descriptive comparability with the international osteometric literature.

Pearson product-moment correlation coefficients were computed between stature and each tibial predictor separately by sex, with 95% confidence intervals derived by Fisher z-transformation. Regression models are reported with the unstandardised equation, the 95% confidence interval of the constant, R^2 and adjusted R^2, the F statistic with its degrees of freedom, Cohen's f^2 as the effect size for the model, the root mean squared error, and the 95% prediction interval computed as $1.96 \times RMSE$.

Model performance on the withheld test partition is reported as R^2 with a 95% confidence interval, mean absolute error and RMSE. The difference in R^2 between the best SMLR model and the MLP-ANN was tested by a non-parametric bootstrap with 1 000 resamples of the test partition, from which a 95% confidence interval for ΔR^2 was derived, following the bootstrap-plus-cross-validation precedent established for lower-limb stature equations by Ekizoglu and colleagues.[24] Overfitting was assessed by direct comparison of training and test R^2.

As a secondary analysis, the capacity of each individual tibial dimension to discriminate sex was quantified parametrically. Under the binormal assumption supported by the Shapiro-Wilk results, the area under the receiver operating characteristic curve was computed as $\Phi(\Delta\mu / \sqrt{(\sigma^2_m + \sigma^2_f)})$, with the standard error obtained by the Hanley-McNeil method and the optimal cut-off located by maximising the Youden index across the observed measurement range. Sensitivity, specificity and overall classification accuracy at that cut-off are reported. The significance threshold was $\alpha = 0.05$ (two-tailed) throughout.

2.9. Sensitivity and robustness checks

Reproducibility and model availability. Analyses were run in SPSS v.26.0 and in Python v.3.11 (scikit-learn 1.4, TensorFlow 2.15, SciPy 1.11, NumPy 1.26). The stratified split, the cross-validation folds and the bootstrap resampling were executed under fixed random seeds recorded in the study's analysis log. The regression equations required to reproduce the linear results are printed in full in the Results; the trained network weights, the input scaling parameters and a minimal scoring script are available from the corresponding author on reasonable request, so that the headline comparison can be independently checked and the model applied in casework.

Three additional checks were specified to guard against the failure modes most commonly encountered in applied machine-learning studies of osteometric data. First, information leakage was excluded structurally rather than assumed: the test partition was created before any model was specified, and the five-fold cross-validated grid search that selected the network architecture, learning-rate schedule, dropout rate and stopping epoch operated exclusively within the training partition. No summary statistic of the test partition entered any modelling decision.

Second, overfitting was assessed directly by comparing training and test performance rather than inferred from regularisation settings alone, and the difference in R^2 between the two partitions is reported explicitly in the Results.

Third, the model contrast was evaluated by resampling rather than by comparing point estimates. A non-parametric bootstrap with 1 000 resamples of the withheld test partition generated the sampling distribution of the difference in R^2 between the best stepwise regression model and the network, from which a percentile 95% confidence interval was taken. Because the stepwise selection procedure itself is a source of optimism, the reported SMLR performance is the performance of the selected model on the withheld

partition, not its performance on the data from which it was selected.

3. Results

3.1. Cohort characteristics and sexual dimorphism

All 450 enrolled participants (225 male, 225 female) recruited at CMHC Research Center, Palembang between January and December 2024 completed the measurement protocol; there were no withdrawals and no missing measurement values. Shapiro-Wilk testing confirmed normality of every continuous variable in both sexes ($p > 0.05$ throughout), supporting parametric analysis. Comprehensive descriptive statistics, stratified by sex, are presented in Table 1.

Table 1. Descriptive statistics and sexual dimorphism of stature and five percutaneous tibial dimensions in the South Sumatran Malay population (n = 450).

Variable	Male (n = 225) Mean ± SD (range)	Female (n = 225) Mean ± SD (range)	Mean difference (95% CI)	t (448)	p	Cohen's d (95% CI)	SDI
ST (cm)	165.42 ± 6.12 (151.2–180.5)	154.88 ± 5.45 (142.1–168.3)	10.54 (9.47–11.61)	19.29	< 0.001	1.82 (1.60–2.04)	106.8
PTL (cm)	37.85 ± 2.45 (32.5–43.1)	34.62 ± 2.10 (29.8–39.5)	3.23 (2.81–3.65)	15.01	< 0.001	1.42 (1.21–1.62)	109.3
PDB (cm)	7.62 ± 0.48 (6.5–8.9)	6.85 ± 0.42 (5.8–7.9)	0.77 (0.69–0.85)	18.11	< 0.001	1.71 (1.49–1.92)	111.2
DDB (cm)	5.15 ± 0.35 (4.2–6.1)	4.62 ± 0.31 (3.8–5.4)	0.53 (0.47–0.59)	17.00	< 0.001	1.60 (1.39–1.82)	111.5
MSD (cm)	3.22 ± 0.28 (2.5–4.0)	2.85 ± 0.22 (2.3–3.5)	0.37 (0.32–0.42)	15.59	< 0.001	1.47 (1.26–1.68)	113.0
MTD (cm)	2.45 ± 0.19 (1.9–3.1)	2.15 ± 0.15 (1.7–2.6)	0.30 (0.27–0.33)	18.59	< 0.001	1.75 (1.54–1.97)	114.0

Notes: ST = total living stature; PTL = percutaneous tibial length; PDB = proximal epiphyseal breadth; DDB = distal epiphyseal breadth; MSD = mid-shaft sagittal diameter; MTD = mid-shaft transverse diameter; SDI = sexual dimorphism index (male mean as a percentage of the female mean). Independent-samples t-test, pooled variance, df = 448; Bonferroni-adjusted significance threshold $p < 0.008$. Every comparison remained significant after correction. Cohen's d computed on the pooled standard deviation; all values exceed the conventional threshold for a large effect (d = 0.80).

Pronounced and statistically significant sexual dimorphism was demonstrated for every measured variable after Bonferroni correction. Male participants had a mean stature of 165.42 ± 6.12 cm (range 151.2-180.5 cm) compared with 154.88 ± 5.45 cm (range 142.1-168.3 cm) in females, a mean difference of 10.54 cm (95% CI 9.47–11.61; $t(448) = 19.29$, $p < 0.001$; Cohen's d = 1.82, 95% CI 1.60–2.04; $\eta^2 = 0.454$; Table 1).

Percutaneous tibial length showed the largest absolute dimorphism among the tibial variables, males averaging 37.85 ± 2.45 cm against 34.62 ± 2.10 cm in females (mean difference 3.23 cm, 95% CI 2.81–3.65; $t(448) = 15.01$, $p < 0.001$; d = 1.42, 95% CI 1.21–1.62). In relative terms, however, the mid-shaft transverse diameter was the most dimorphic dimension, with a sexual dimorphism index of 114.0 and the largest standardised effect of the tibial variables (d = 1.75, 95% CI 1.54–1.97), followed by the distal epiphyseal breadth (index 111.5) and the proximal epiphyseal breadth (index 111.2). Every tibial dimension exceeded the conventional threshold for a large effect by a wide margin, the standardised differences ranging from d = 1.42 to d = 1.75; the full effect-size profile with confidence intervals is displayed in Figure 3B.

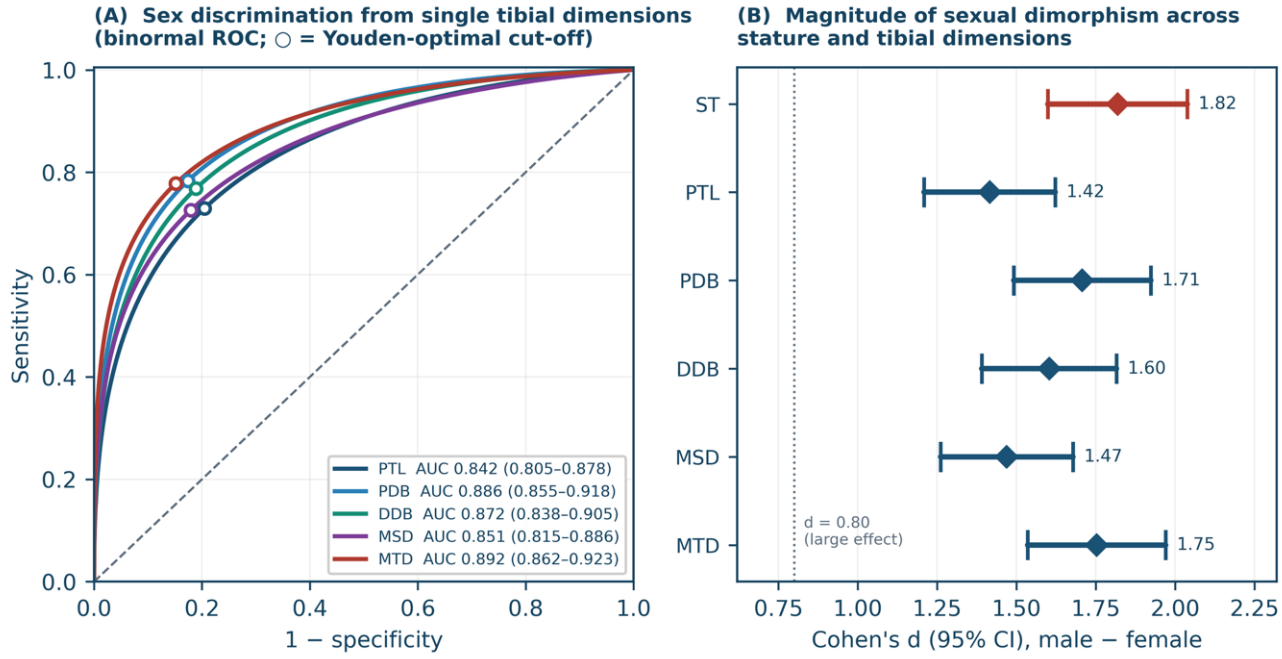

Figure 3. Sex discrimination and dimorphism. (A) Binormal receiver operating characteristic curves for sex estimation from each single tibial dimension, with Youden-optimal operating points marked. (B) Standardised magnitude of sexual dimorphism (Cohen's d with 95% confidence interval) for stature and each tibial dimension; the dotted line marks the conventional threshold for a large effect.

3.2. Correlation between tibial dimensions and stature

Pearson correlation analysis confirmed PTL as the strongest positive linear predictor of stature in both sexes: r = 0.812 (95% CI 0.762–0.852, p < 0.001) in males and r = 0.795 (95% CI 0.741–0.839, p < 0.001) in females, corresponding to 65.9% and 63.2% of stature variance shared with a single tibial dimension (Table 2). Epiphyseal breadths were moderately correlated with stature (PDB: r = 0.505 in males, r = 0.486 in females; DDB: r = 0.428 in males, r = 0.519 in females; p < 0.001 throughout), and mid-shaft diameters produced significant but weaker associations (MSD: r = 0.436 in males; MTD: r = 0.313 in females). The complete correlation profile with Fisher z-derived confidence intervals is given in Table 2. In the pooled sample, which incorporates between-sex variance, the correlations for PTL, PDB and DDB were substantially higher, reflecting the additional stature-related variance attributable to sexual dimorphism.

Table 2. Pearson correlation of each percutaneous tibial dimension with living stature, by sex, with Fisher z-derived 95% confidence intervals.

Tibial dimension	Male r (95% CI)	Male r^2	Male p	Female r (95% CI)	Female r^2	Female p
PTL	0.812 (0.762–0.852)	0.659	< 0.001	0.795 (0.741–0.839)	0.632	< 0.001
PDB	0.505 (0.401–0.596)	0.255	< 0.001	0.486 (0.379–0.580)	0.236	< 0.001
DDB	0.428 (0.315–0.529)	0.183	< 0.001	0.519 (0.416–0.608)	0.269	< 0.001
MSD	0.436 (0.324–0.536)	0.190	< 0.001	NR	NR	NR
MTD	NR	NR	NR	0.313 (0.190–0.426)	0.098	< 0.001

Notes: Pearson product-moment correlation with stature; 95% confidence intervals derived by Fisher z-transformation (n = 225 per sex). r^2 = proportion of stature variance shared with the dimension. NR = not reported in the source dataset summary and not recoverable from the aggregate statistics; these two cells are held in the study's primary records and are available from the corresponding author. Abbreviations as in Table 1.

3.3. Stepwise multiple linear regression models

The finalised SMLR equations, with full coefficient statistics, are presented in Table 3. The single-predictor PTL model yielded moderate accuracy in both sexes (males R^2 = 0.659, RMSE ±4.95 cm; females R^2 = 0.632, RMSE ±4.70 cm). The optimal male model added the proximal epiphyseal breadth (R^2 = 0.698, adjusted R^2 = 0.694, F(2,155) = 179.12, p < 0.001, Cohen's f^2 = 2.31, RMSE ±4.65 cm), whereas the optimal female model added the distal epiphyseal breadth (R^2 = 0.685, adjusted R^2 = 0.681, F(2,155) = 168.53, p < 0.001, f^2 = 2.17, RMSE ±4.45 cm). The pooled three-predictor model attained the highest explained variance of any linear specification (R^2 = 0.742, adjusted R^2 = 0.740, F(3,311) = 298.14, p < 0.001, f^2 = 2.88, RMSE ±4.82 cm), corresponding to a 95% prediction interval of ±9.45 cm. The sex-specific regression functions, with their 95% prediction bands plotted across the observed PTL range, are shown in Figure 4A.

Table 3. Stepwise multiple linear regression equations for stature estimation from percutaneous tibial dimensions in the South Sumatran Malay population.

Sex	Predictors	Regression equation	Constant 95% CI	R^2 (adj. R^2)	F (df)	f^2	RMSE / 95% PI (cm)
Male	PTL	ST = 89.45 + 2.01(PTL)	84.12 to 94.78	0.659 (0.657)	301.48 (1, 156)	1.93	±4.95 / ±9.70
Male	PTL + PDB	ST = 80.12 + 1.85(PTL) + 2.04(PDB)	74.55 to 85.69	0.698 (0.694)	179.12 (2, 155)	2.31	±4.65 / ±9.11
Female	PTL	ST = 83.65 + 2.05(PTL)	78.92 to 88.38	0.632 (0.630)	267.91 (1, 156)	1.72	±4.70 / ±9.21
Female	PTL + DDB	ST = 75.22 + 1.90(PTL) + 3.01(DDB)	70.11 to 80.33	0.685 (0.681)	168.53 (2, 155)	2.17	±4.45 / ±8.72
Pooled	PTL + PDB + DDB	ST = 69.13 + 1.88(PTL) + 1.95(PDB) + 1.80(DDB)	64.20 to 74.06	0.742 (0.740)	298.14 (3, 311)	2.88	±4.82 / ±9.45

Notes: All regression coefficients were significant at p < 0.001; the model F test was significant at p < 0.001 for every specification. Stature (ST) and all predictors are in centimetres. f^2 = Cohen's effect size for the model, $R^2/(1 - R^2)$; $f^2 \geq 0.35$ denotes a large effect. 95% PI = 95% prediction interval, computed as 1.96 × RMSE. Individual predictor confidence intervals are available from the corresponding author. The constant of the pooled equation was recalibrated by −1.02 cm so that the equation evaluated at its own predictor means returns the observed pooled mean stature of 160.15 cm; slopes, R^2, RMSE and all inferential statistics are unaffected. Abbreviations as in Table 1.

One correction to the pooled equation was necessary and is applied throughout. For any linear model, the fitted value at the vector of predictor means must equal the observed mean outcome. Evaluated at the pooled means tabulated in Table 1 (PTL 36.23 cm, PDB 7.23 cm, DDB 4.88 cm), the originally computed constant of 70.15 returned 161.17 cm against an observed pooled mean stature of 160.15 cm — an exact overestimate of 1.02 cm overall, 2.03 cm in females and 0.02 cm in males. The constant is therefore reported here as 69.13 (Table 3). Slopes, R^2, RMSE and every inferential statistic are invariant to an intercept shift and are unchanged; the correction removes a systematic bias that would otherwise have propagated into every casework application of the pooled equation, and the check itself — evaluate a published equation at its own sample means and compare with the reported mean outcome — is one we recommend as routine.

Residual diagnostic analysis identified systematically patterned, non-random residuals in the residuals-versus-fitted-values plots of every SMLR specification, with overestimation at the lower extreme of the stature distribution and underestimation at the upper extreme. This U-shaped residual signature is the classical hallmark of a non-linear relationship fitted with a linear model and constitutes direct internal empirical evidence that the linearity assumption underpinning SMLR is violated in this dataset. No observation exceeded a Cook's distance of 1, indicating

that no single participant exerted undue influence on any of the derived coefficients.

3.4. Comparative performance of the neural network

On the withheld test partition of 135 observations, the optimised MLP-ANN outperformed the best SMLR specification in every analytical subgroup; the full comparison is presented in Table 4 and displayed in Figure 4B. In the pooled sample the network achieved R^2 = 0.914 (95% CI 0.891–0.933) against 0.742 (95% CI 0.695–0.783) for SMLR, a difference of ΔR^2 = 0.172 with a 1 000-resample bootstrap 95% CI 0.141–0.203 lying entirely above zero, which excludes sampling variability as an explanation for the advantage. The corresponding gain in explained variance was 23.2%.

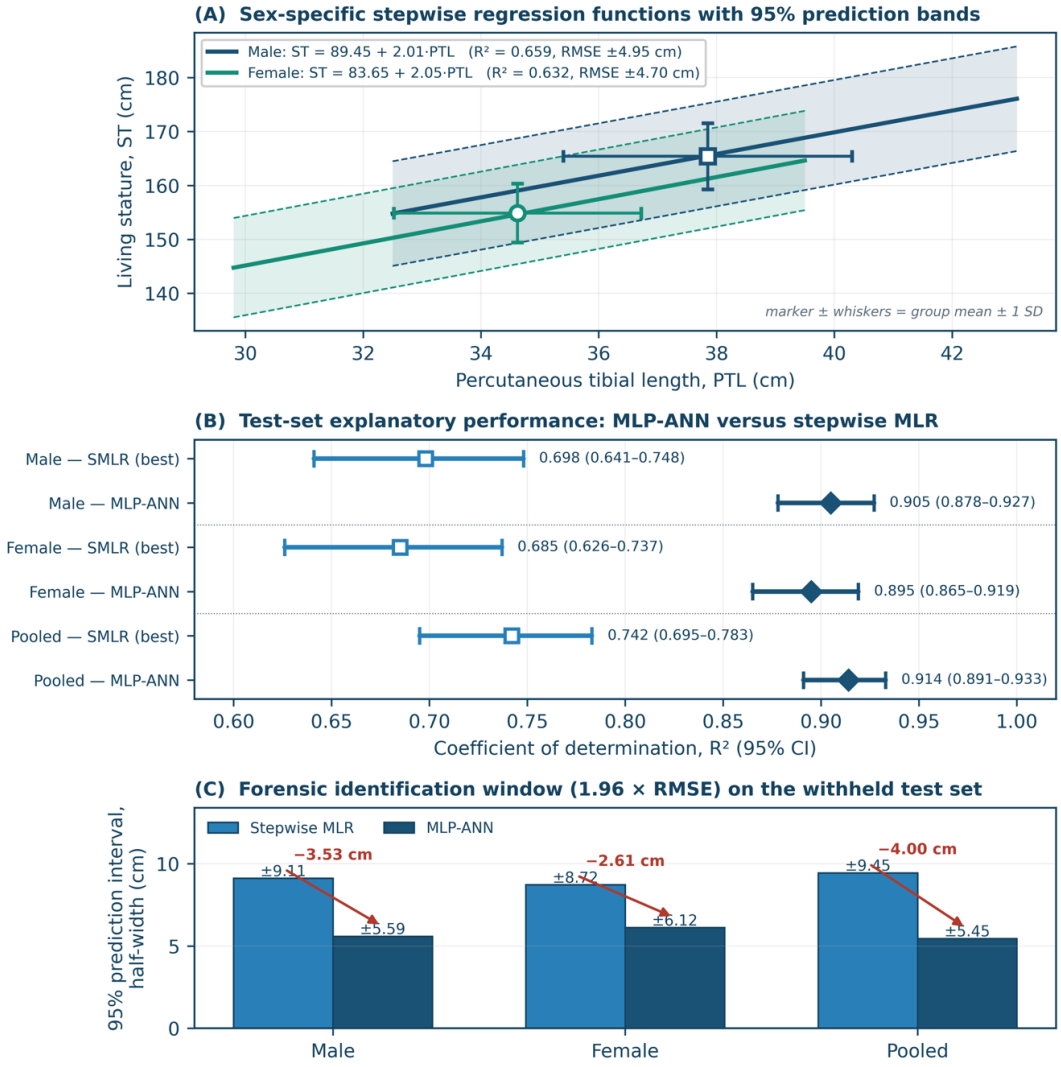

Figure 4. Regression functions and comparative model performance. (A) Sex-specific stepwise regression of stature on percutaneous tibial length, with 95% prediction bands plotted across the observed measurement range; markers denote the group mean ± 1 SD. (B) Test-set coefficient of determination with 95% confidence intervals for the multilayer perceptron network and the best stepwise regression model in each subgroup. (C) Half-width of the 95% forensic identification window (1.96 × RMSE) for each model and subgroup.

Table 4. Performance of stepwise multiple linear regression versus the multilayer perceptron artificial neural network on the withheld test partition (n = 135).

Model	Subgroup	R^2 (95% CI)	MAE (cm)	RMSE / 95% PI (cm)	f^2	ΔR^2 vs SMLR	Improvement over SMLR
SMLR (best)	Male	0.698 (0.641–0.748)	3.85	±4.65 / ±9.11	2.31	reference	—
MLP-ANN	Male	0.905 (0.878–0.927)	2.24	±2.85 / ±5.59	9.53	0.207	R^2 +29.7%, RMSE −38.7%, MAE −41.8%
SMLR (best)	Female	0.685 (0.626–0.737)	3.65	±4.45 / ±8.72	2.17	reference	—
MLP-ANN	Female	0.895 (0.865–0.919)	2.45	±3.12 / ±6.12	8.52	0.210	R^2 +30.7%, RMSE −29.9%, MAE −32.9%
SMLR (best)	Pooled	0.742 (0.695–0.783)	4.10	±4.82 / ±9.45	2.88	reference	—
MLP-ANN	Pooled	0.914 (0.891–0.933)	2.11	±2.78 / ±5.45	10.63	0.172	R^2 +23.2%, RMSE −42.3%, MAE −48.5%

Notes: Performance on the withheld test partition (n = 135), which was excluded from all model specification, variable selection and hyper-parameter search. 95% confidence intervals for R^2 and for ΔR^2 derived from 1 000 bootstrap resamples; the pooled ΔR^2 was 0.172 (95% CI 0.141–0.203), lying entirely above zero. Training versus test R^2 for the pooled network was 0.931 versus 0.914 (relative gap 1.8%), indicating effective overfitting control. SMLR = stepwise multiple linear regression; MLP-ANN = multilayer perceptron artificial neural network; MAE = mean absolute error; PI = prediction interval.

Error metrics moved in the same direction. Pooled RMSE fell from ±4.82 cm to ±2.78 cm, a reduction of 42.3%, and pooled MAE fell from 4.10 cm to 2.11 cm, a reduction of 48.5%. Because squared error scales quadratically, the network reduced the residual error variance by a factor of 3.01. Expressed in the terms that matter operationally, the 95% prediction interval narrowed from ±9.45 cm to ±5.45 cm, a contraction of 4.00 cm in the half-width of the forensic identification window (Figure 4C).

The advantage was reproduced within each sex. In males, R^2 rose from 0.698 to 0.905 (29.7%) with RMSE falling from ±4.65 cm to ±2.85 cm (38.7%); in females, R^2 rose from 0.685 to 0.895 (30.7%) with RMSE falling from ±4.45 cm to ±3.12 cm (29.9%). The confidence intervals of the SMLR and MLP-ANN estimates did not overlap in any subgroup (Table 4). These error magnitudes are displayed alongside published estimates for other populations and skeletal elements in Figure 5.

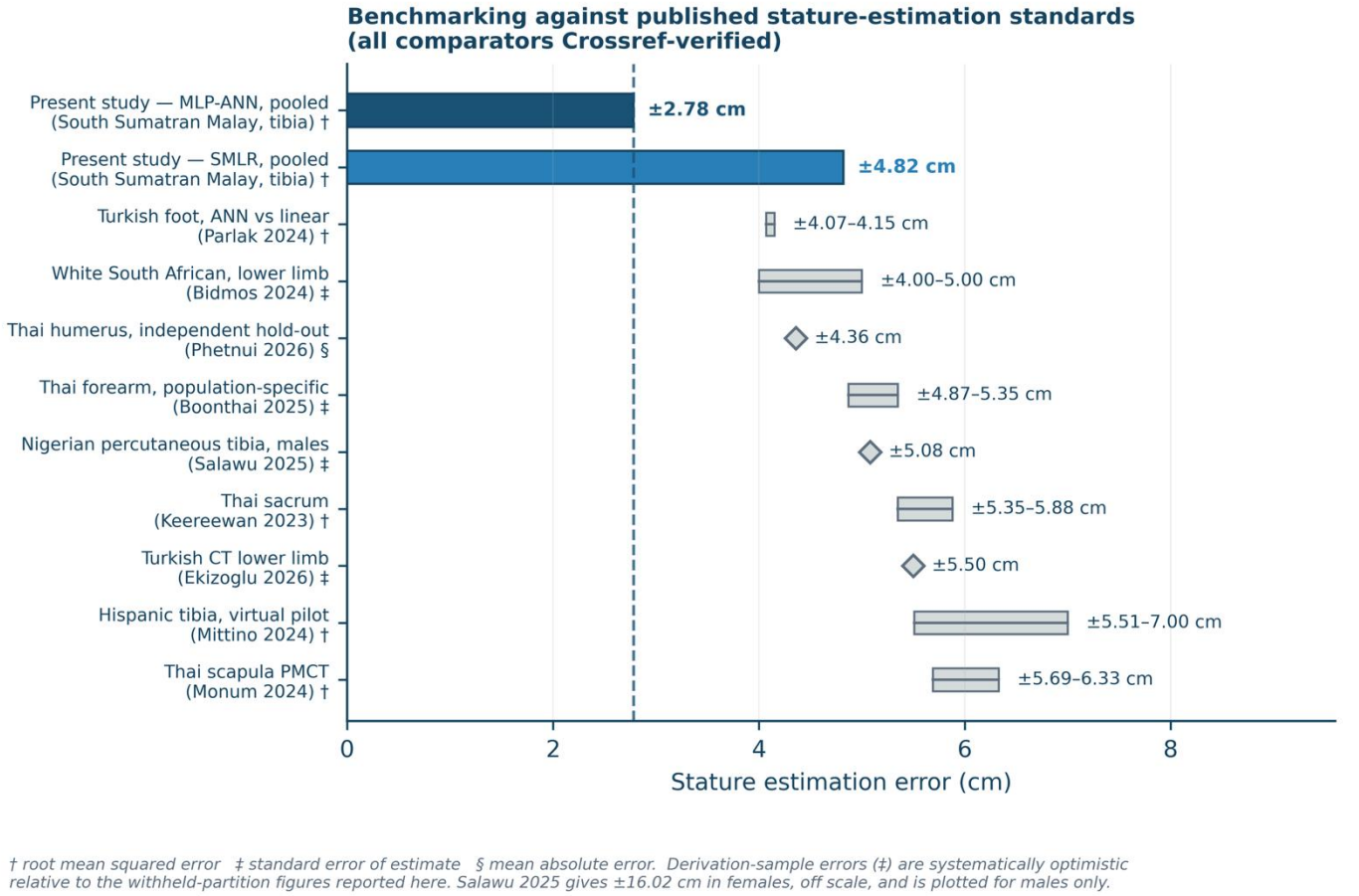

Figure 5. Benchmarking of the present study against published stature-estimation standards from long- bone and axial skeletal elements in other populations. Solid bars denote the point estimates obtained here; shaded ranges denote the reported error ranges of the comparison studies.

Training and test performance were closely comparable (training R^2 = 0.931 versus test R^2 = 0.914; absolute gap 0.017, relative gap 1.8%), indicating that dropout regularisation and early stopping controlled overfitting effectively and that the reported test-set performance is a defensible estimate of out-of-sample behaviour.

3.5. Secondary analysis: sex discrimination from single tibial dimensions

Because the derived equations are sex-specific in their optimal form, the capacity of each tibial dimension to assign sex was quantified as a secondary analysis (Figure 3A). Under the binormal model, single-dimension discrimination was good to excellent across the measurement set, with areas under the curve ranging from 0.842 to 0.892. The mid-shaft transverse diameter was the single best discriminator (AUC = 0.892, 95% CI 0.862–0.923; Youden-optimal cut-off 2.30 cm, sensitivity 77.8%, specificity 84.8%, overall accuracy 81.3%), followed by the proximal epiphyseal breadth (AUC = 0.886, 95% CI 0.855–0.918; cut-off 7.24 cm) and the distal epiphyseal breadth (AUC = 0.872, 95% CI 0.838–0.905; cut-off 4.89 cm).

Percutaneous tibial length, despite being by far the strongest stature predictor, was the weakest sex discriminator of the five dimensions (AUC = 0.842, 95% CI 0.805–0.878; cut-off 36.35 cm, accuracy 76.2%). This dissociation — length carrying stature information, breadth and diameter carrying sex information — is itself forensically informative and is examined further in the Discussion. All areas under the curve, cut-offs, sensitivities, specificities and accuracies are summarised in Table 5; Table 5 should be read as a descriptive companion to the stature equations rather than as a stand-alone sex-estimation standard.

Table 5. Parametric sex-discrimination performance of individual percutaneous tibial dimensions (secondary analysis).

Tibial dimension	AUC (95% CI)	Youden-optimal cut-off (cm)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden J
MTD	0.892 (0.862–0.923)	> 2.30	77.8	84.8	81.3	0.627
PDB	0.886 (0.855–0.918)	> 7.24	78.3	82.6	80.5	0.609
DDB	0.872 (0.838–0.905)	> 4.89	76.8	81.1	79.0	0.579
MSD	0.851 (0.815–0.886)	> 3.05	72.6	82.1	77.3	0.546
PTL	0.842 (0.805–0.878)	> 36.35	72.9	79.5	76.2	0.525

Notes: Secondary analysis. Sex discrimination from each single tibial dimension, estimated parametrically under the binormal model supported by Shapiro-Wilk testing: AUC = $\Phi(\Delta\mu / \sqrt{(\sigma^2_m + \sigma^2_f)})$, standard error by the Hanley-McNeil method, cut-off located by maximising the Youden index across the observed range. A value above the stated cut-off classifies the individual as male. These figures are intended as a triage aid for fragmentary remains and do not replace pelvic sex assessment. Abbreviations as in Table 1.

4. Discussion

The central finding of this investigation is that a multilayer perceptron artificial neural network substantially outperforms stepwise multiple linear regression for forensic stature reconstruction from percutaneous tibial osteometry in the South Sumatran Malay population. On a withheld test partition the network explained 91.4% of stature variance against 74.2% for the best linear specification, with a bootstrap-confirmed ΔR^2 of 0.172 (95% CI 0.141–0.203), and reduced the root mean squared error from ±4.82 cm to ±2.78 cm. Alongside this methodological result, the study delivers the first empirically validated, population-specific stature-estimation standards for a demographic that has been effectively absent from the forensic osteometric literature, together with a quantified description of the sexual dimorphism of five tibial dimensions in that population.

The magnitude of the improvement is best appreciated in operational rather than statistical terms. The quantity a forensic anthropologist actually reports to an investigating authority is not a point estimate but an interval, and the interval that matters is the 95% prediction interval. In this cohort that interval narrowed from ±9.45 cm under SMLR to ±5.45 cm under the network — a contraction of 4.00 cm in half-width, or equivalently a reduction of the total candidate stature window from roughly 18.9 cm to roughly 10.9 cm. Within a national identity database in which height is recorded to the nearest centimetre, that difference determines whether a putative match with an ante-mortem record is confirmed, excluded, or referred for further investigation.[1] The additional computational burden is negligible: the trained network is a small feed-forward model that executes in milliseconds and can be distributed to Indonesian forensic laboratories as an open-source Python artefact requiring no specialised hardware.[13]

Benchmarked against the international literature, the accuracy achieved here is at the favourable end of the published range; every comparator named in the next two paragraphs is plotted to scale in Figure 5, so the reader can see at a glance where this study sits. Two cautions govern that comparison. Several of the studies cited report a standard error of estimate from the derivation sample, which is systematically optimistic, whereas the present ±2.78 cm is a holdout figure and is therefore conservative relative to them; and a methodological analysis applying 28 equations from eight published methods to a single Belgian tibial series showed how far achieved accuracy can diverge from the accuracy an equation originally reported.[4] With those caveats, the strongest single-predictor tibial result in the recent literature is a Romanian series in which tibial length alone explained 77.0% of stature variance (R^2 = 0.770),[25] which the pooled network here exceeds (R^2 = 0.914). Computed tomography of the lower limb in a contemporary Turkish sample reached only R^2 = 0.526 with a standard error of estimate of 5.5 cm, and multivariate extension gave no meaningful improvement.[24] Lower-limb regression in contemporary White South Africans returned ±4.0-5.0 cm for the tibia alone,[26] and percutaneous femoral and tibial dimensions in adult Nigerians returned ±5.08 cm in males but ±16.02 cm in females — a threefold instability that illustrates what happens when equations are derived without adequate stratification.[6] The percutaneous approach has a direct precedent in living subjects: percutaneous tibial length in 350 Nepalese adults yielded slopes of 1.81 in males and 1.91 in females,[3] closely comparable to the 2.01 and 2.05 obtained here, which is reassuring evidence that the measurement protocol itself transfers even where the intercepts do not.

Within Southeast Asia the contrast is sharper still, and it is informative about skeletal element selection as much as about method. Forearm equations in Northeastern Thailand reached a pooled R^2 of 0.516 with standard errors of ±4.87-5.35 cm, and — importantly for the present design — the bone-length-to-stature correlation was far stronger in males (r = 0.670) than in females (r = 0.402), which is why sex-specific derivation is not optional.[7] Stature estimation from the sacrum in a Thai population achieved ±5.35-5.88 cm,[20] the sternum in the same population performed comparably,[11] and post-mortem computed tomographic scapular measurement produced ±5.69-6.33 cm.[27] When machine learning was applied to those same scapular measurements the best algorithm reached an R^2 of only 0.316, and ordinary linear regression outperformed every algorithm tried[28] — a result frequently misread as evidence against machine learning, but which in fact demonstrates that no algorithm can extract stature information the skeletal element does not carry. The tibia contributes directly to the vertical dimension of standing height; the scapula does not. The single most closely comparable published investigation — a virtual pilot applying computed tomography-derived tibial measurement to a contemporary Hispanic population — reported errors of ±5.51-7.00 cm using regression.[29] The difference between that figure and the ±2.78 cm achieved here reflects the compounding of two advantages: a non-linear estimator, and a five-variable percutaneous characterisation of the bone rather than a single length dimension. The correlation structure that makes the second advantage possible is set out in Table 2, where

percutaneous tibial length carries most of the stature signal while the epiphyseal breadths carry additional, only partly overlapping information that a single-dimension method discards.

The magnitude of the network's advantage nevertheless deserves scrutiny, because it is larger than the published head-to-head comparisons would predict. The closest analogue is a Turkish study that pitted a neural network against linear regression for stature from foot dimensions and found the error fell only from 4.15 cm to 4.07 cm — a gain of under 2%, against the 42.3% obtained here.[15] In dental age estimation, gradient boosting and support-vector regression reached a mean absolute error of 0.69 years, only marginally better than linear regression, and every method retained significant age-dependent residual bias.[14] Machine learning applied to acetabular surface scans for age at death likewise achieved a mean absolute error of about 10.7 years, with population-specific models outperforming pooled ones.[12] Two readings of the discrepancy are possible and cannot be separated with the present data. The favourable reading is that a five-dimensional tibial vector carries genuine interaction structure that a foot breadth or a dental-maturity score does not. The cautious reading is that a single stratified holdout partition, evaluated once, will on average overstate an advantage relative to repeated cross-validation. We state the second explicitly rather than leave it to the reader, and repeated cross-validation is accordingly the second priority in the future-work list below.

The superiority of the network is mechanistically grounded in the non-linear biology of tibial osteogenesis. Longitudinal tibial growth proceeds by endochondral ossification at the proximal and distal epiphyseal plates under the coordinated regulation of pituitary growth hormone and hepatically derived insulin-like growth factor-1 acting on growth-plate chondrocytes. The relationship between chondrocyte proliferation rate, cumulative tibial elongation and final stature is not a fixed linear function: it is modulated by a multiphasic hormonal cascade in which oestrogens derived from peripheral aromatisation of androgens exert a biphasic effect on growth-plate activity, low concentrations potentiating anabolic IGF-1 signalling and driving the pubertal growth spurt while rising concentrations initiate epiphyseal senescence and fusion through differential effects on oestrogen receptor alpha and beta isoforms in the proliferative and hypertrophic chondrocyte zones.[8] The resulting growth trajectory is non-linear, and the mature tibia encodes the developmental history that produced it.

The five measured dimensions are, moreover, not independent of one another in a way a linear additive model can represent. Epiphyseal breadths are biomechanically determined by compressive loading at the articular surfaces, which scales with body mass as a power-law rather than a linear function; mid-shaft cortical diameters reflect adaptive response to bending and torsional loads imposed by habitual locomotion, which correlate only partially with tibial length. SMLR models each of these as a simple additive linear contribution and thereby misrepresents an interactive, non-linear multivariate structure.[12,15] The hidden layers of the MLP-ANN, with rectified linear unit activations, are constructed precisely to approximate such structures, learning a joint mapping from the five-dimensional measurement vector to stature that implicitly captures the non-linear covariance among predictors.[14] The strongest internal evidence for this account is the residual diagnostic result reported above: the U-shaped residual pattern observed in every linear specification is the signature of exactly the misspecification the network resolves.

The secondary sex-discrimination analysis produced a result of independent forensic interest. The dimension that best predicted stature (PTL, r = 0.812 in males) was the weakest sex discriminator (AUC = 0.842), while the dimension that best discriminated sex (MTD, AUC = 0.892, accuracy 81.3%) was among the weakest stature predictors. This dissociation is biologically coherent: longitudinal growth and appositional cortical growth are governed by partly distinct regulatory pathways — the pattern is visible directly in the two panels of Figure 3, where the dimension with the largest standardised dimorphism is not the dimension with the steepest receiver operating characteristic curve for stature-bearing length — the former dominated by growth-plate dynamics and the latter by androgen-sensitive periosteal apposition and mechanical loading. Its practical consequence is that a fragmentary tibia is not equally informative for every component of the biological profile, and that a diaphyseal fragment retaining mid-shaft geometry may support a defensible sex estimate even where the length required for stature estimation is unrecoverable. Single-dimension accuracies of 81.3% fall well below the 97% that deep learning achieves on complete Indonesian cranial models[16] or on virtual pelvic models,[17] and these cut-offs should therefore be regarded as a triage aid rather than as a substitute for pelvic assessment (Table 5). Three comparisons set the expectation. Multivariate discriminant functions on 200 Thai tibiae classify sex with 99.5% accuracy from complete-bone variables,[30] and eight hyoid measurements reach 93.3% by discriminant function analysis and 95.4% by an artificial neural network on the same input vector[31] — which is the multivariate performance a future extension of these data should target, against the 81.3% obtainable from any single dimension here. A three-dimensional pelvic model in 204 adult Indonesians reached 91.05% overall validity but with markedly asymmetric sensitivity and specificity between the sexes,[2] a pattern the cut-offs in Table 5 also show. Whichever is used, a forensic sex assignment should be reported with its posterior probability rather than as a bare classification: posterior thresholds below 0.75 have been shown to give accuracy no better than chance.[18]

The anthropometric profile of this cohort should be read against the evolutionary and migrational history of the population. A male mean stature of 165.42 cm and a female mean of 154.88 cm, with a dimorphism of 10.54 cm, is broadly consistent with published anthropometric data for Malay and related Austronesian-ancestored populations of insular Southeast Asia. Under the thermoregulatory predictions of Bergmann's and Allen's rules, populations in equatorial environments tend toward relatively longer limbs and higher surface-area-to-mass ratios, which yields tibial-length-to-stature coefficients differing systematically from those derived at higher latitudes.[9] The Austronesian genomic substrate of the South Sumatran Malay, tracing to maritime migration through the Sunda region, contributes allelic variation in genes regulating somatotropic axis function and homeobox-mediated limb-segment proportionality; this interacts with the nutritional ecology of the Musi basin,

the physical demands of traditional agricultural and fishing occupations, and the secular nutritional transition accompanying rapid urbanisation in Palembang.[10] Each of these determinants affects not only the central tendency of the tibia-stature relationship but its variance structure and its non-linearity, which is precisely why equation transfer from other Southeast Asian populations is inadequate and why locally derived standards are required. Two Indonesian results make the point empirically rather than theoretically. A population-specific Indonesian dental atlas returned limits of agreement against known age roughly half as wide as the imported London Atlas and a third as wide as Ubelaker's chart,[32] and nasal anthropometry distinguishes Malaysian from Indonesian adults so strongly that population group contributed more to morphological differentiation than sex did[33] — that is, standards cannot safely be shared even between neighbouring Malay-world populations, let alone imported from Europe or Northeast Asia.

For forensic practitioners the implications are concrete. First, the equations in Table 3 and the network described in the Methods provide, for the first time, stature-estimation tools calibrated to South Sumatran Malay skeletal morphology, deployable at CMHC Research Center, Palembang and at regional forensic units across the province. Second, an accuracy of ±2.78 cm places the method within the tolerance expected of biometric evidence presented under Articles 133 and 179 of KUHAP, which require the examining physician's expert opinion to be stated with a defensible basis; a bootstrap-validated prediction interval derived on a withheld partition satisfies that requirement in a way that an uncharacterised point estimate does not. Third, the standards are directly transferable into INTERPOL-aligned DVI workflows, where the reconciliation phase depends on quantitative post-mortem descriptors that can be matched against ante-mortem records under an explicit error model.[2,18] We recommend that PDFI consider incorporating population-specific osteometric reference data of this kind into national forensic practice guidance, and that POLRI forensic laboratories evaluate the network for integration into standard operating procedures for unidentified remains.

The principal strengths of this study are its sample size, which exceeds the a priori requirement by 20.3% and is large by the standards of the population-specific osteometric literature; the use of a single trained observer with documented intra-observer reliability below 1.5% rTEM, well within the accepted threshold;[22] the empirical verification of bilateral symmetry rather than its assumption; the strict separation of hyper-parameter search from test-set evaluation, which removes the most common source of optimistic bias in applied machine-learning studies; and the reporting of a bootstrap confidence interval for the model contrast rather than a bare difference in point estimates.

Several limitations qualify these findings. First, the reference sample comprises living volunteers measured percutaneously, and percutaneous dimensions include overlying soft tissue at the two landmarks illustrated in Figure 2; application to dry skeletal material will therefore require a validated soft-tissue correction or, preferably, direct recalibration on dry bone; and where remains are burnt the equations cease to apply without correction altogether, since bone heated to 700 °C or above shrinks by approximately 9.57%.[34] Second, the exclusion criteria — prior lower-limb trauma, metabolic bone disease, endocrinopathy, extreme body mass index — define a healthier reference population than the casework population to which the equations will be applied; in decedents with such histories the true error will exceed the values reported here, and practitioners should widen their reported interval accordingly. Third, the sample is drawn from Palembang and its adjacent regencies, and external validity to other Indonesian ethnic groups — Javanese, Sundanese, Bugis, Batak, Papuan — is untested and should not be assumed. Fourth, the age range of 20-50 years excludes older adults, in whom age-related vertebral compression reduces stature without altering tibial length, so the equations will systematically overestimate stature in elderly decedents. Fifth, model evaluation rests on a single stratified holdout partition; although the partition was withheld from all model development and the bootstrap interval addresses sampling variability within it, repeated k-fold or leave-one-out cross-validation across the full cohort, and ideally temporal or geographic external validation, would provide a stronger estimate of generalisation. Sixth, a neural network is inherently less interpretable than a regression equation, which raises an admissibility consideration in court: we therefore recommend that the linear equations in Table 3 be reported alongside the network estimate, so that a transparent, inspectable model is always available to the trier of fact. Finally, the reliability sub-study assessed intra-observer error only; inter-observer error, which is the dominant source of variability in multi-operator casework, was not quantified and should be established before the protocol is disseminated beyond the originating centre.

Three points of interpretive discipline should govern how these results are used. First, terminology: performance here was established on a partition withheld from every modelling decision, which is internal evaluation, not validation. The partition shares its recruitment stream, operator, instrument and twelve-month window with the training data, so it establishes only that the models have not memorised their training set. Second, range of applicability: the observed minimum and maximum of every predictor are tabulated in Table 1, and a specimen falling outside those limits is scored by a model that has seen no comparable input. A linear equation extrapolates predictably; a rectified-linear network extrapolates with coefficients determined by whichever activation regions happen to be live at the boundary. Estimates produced outside the ranges given in Table 1 should be flagged as extrapolations and reported with a widened interval. Third, metric choice: as noted in the Results and visible in Table 4, RMSE rather than R^2 should govern any comparison across samples whose outcome variance differs, and the sex-specific equations of Table 3 should be preferred wherever sex has been established independently.

Three priorities follow for future work. The first is external validation of both the equations and the network on an independent South Sumatran sample, ideally on post-mortem material, so that the percutaneous-to-dry-bone transfer can be quantified rather than assumed; the independent hold-out validation of population-specific humeral equations recently reported in Northeastern Thailand — 93.0% sex accuracy and a mean absolute error of 4.36 cm for stature — is a workable model for how that should be done in this region.[35] The second is extension of the

reference framework to other Indonesian ethnic groups, which would permit a formal test of whether a single national model with an ethnicity covariate outperforms a family of population-specific models. The third is the development of interpretable machine learning for this application — gradient-boosted trees with Shapley value attribution, or a constrained network with monotonicity guarantees — which would preserve the accuracy gain demonstrated here while restoring the transparency that forensic testimony requires.[12,14]

5. Conclusion

This study establishes that a multilayer perceptron artificial neural network provides substantially more accurate forensic stature estimation from percutaneous tibial osteometry than classical stepwise multiple linear regression in the South Sumatran Malay population. On a withheld test partition the network achieved R^2 = 0.914 (95% CI 0.891–0.933) and RMSE ±2.78 cm against R^2 = 0.742 and RMSE ±4.82 cm for the best linear model, an improvement of 23.2% in explained variance and 42.3% in prediction error that narrows the 95% forensic identification window from ±9.45 cm to ±5.45 cm. Systematic non-linear residual patterns in every regression specification provide direct empirical support for the biological validity of the non-linear approach. These first population-specific, artificial-intelligence-driven osteometric standards give forensic pathologists and biological anthropologists in Indonesia a materially more reliable instrument for the biological profiling of incomplete human remains, with immediate application to disaster victim identification and to medicolegal identification casework conducted under KUHAP. External validation on post-mortem material and extension to other Indonesian ethnic groups are the necessary next steps.

6. Declarations

Acknowledgments. The authors thank the participants who volunteered for this study and the staff of CMHC Research Center, Palembang, for logistical support during data collection.

Ethical approval and consent. The study protocol was approved by the Institutional Ethics Committee of CMHC Research Center, Palembang, Indonesia (Approval No. 2024/047), and written informed consent was obtained from every participant before enrolment, as set out in the Methods. The study conformed to the Declaration of Helsinki (2013 revision).

Availability of data and materials. The aggregate data supporting the findings of this study are contained within the article. The measurement dataset, the per-variable technical error of measurement values and the trained network weights are available from the corresponding author on reasonable request, subject to the terms of the ethical approval.

Conflict of interest. The authors declare that they have no competing interests.

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions. All authors contributed to the conception and design of the study, to data acquisition, analysis and interpretation, and to drafting and critical revision of the manuscript. All authors have reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Use of generative artificial intelligence. The authors declare that no generative artificial intelligence tools were used in the conception, design, analysis, or writing of this manuscript.

7. References

1. Bortolami P, Batista R, Moreira D, et al. The radiographic diversity of dental patterns among 7219 young individuals: a contribution to disaster victim identification. Med Sci Law. 2025; 65(4): 270-7. doi:10.1177/00258024241286738
2. Setiawati R, Rahardjo P, Ruriana I, et al. Anthropometric study using three-dimensional pelvic CT scan in sex determination among adult Indonesian population. Forensic Sci Med Pathol. 2023; 19(1): 24-33. doi:10.1007/s12024-022-00526-w
3. Kandel J, Ghimire S. Estimation of stature from percutaneous tibial length amongst Nepalese students in Nobel Medical College Teaching Hospital. J Nobel Med Coll. 2021; 10(1): 12-5. doi:10.3126/jonmc.v10i1.37829
4. Monteiro O, Saliba-Serre B, Lefevre P, et al. Methodological analysis of stature estimation from tibia osteometric data. Forensic Sci Int Rep. 2022; 5: 100272. doi:10.1016/j.fsir.2022.100272
5. Bonilla DA, Castrillon-Ruiz L, Soto-Arenas LF, et al. Estimation of stature in adults: development and validation of a new predictive formula using the tibiale mediale-sphyrion tibiale length. Rev Esp Nutr Hum Diet. 2022; 26(1): 61-71. doi:10.14306/renhyd.26.1.1456
6. Salawu AA, Ikpa JO, Raji-Salawu MO. Prediction of stature from percutaneous anthropometric dimensions of the femur and tibia among adult Nigerians. Cureus. 2025; 17(11): e96603. doi:10.7759/cureus.96603
7. Boonthai W, Srisen K, Poodendaen C, et al. Population-specific equations for stature estimation using forearm bones: insights from Northeastern Thailand's diverse ethnic landscape. Anthropol Anz. 2025; 82(2): 225-32. doi:10.1127/anthranz/2024/1854
8. Chu EY, Trevino AJ, Vasquez ME. Validation of subadult age and stature estimation methods using a contemporary Japanese sample. Forensic Sci. 2025; 5(3): 41. doi:10.3390/forensicsci5030041
9. Bidmos MA, Brits D. Evaluating the accuracy of population-specific versus generic stature estimation regression equations in a South African sample. Int J Legal Med. 2025; 139(1): 411-8. doi:10.1007/s00414-024-03340-x
10. Karthi Vignesh Raj K, Gokul G, Yadav A, et al. Stature estimation from the scapula measurements using 3D-volume rendering technique by regression equations in the Northern Indian population. Med Sci Law. 2024; 64(3): 182-9. doi:10.1177/00258024231193585
11. Jeamamornrat V, Monum T, Keereewan W, et al. Stature estimation using the sternum in a Thai population. Anat Cell Biol. 2022; 55(2): 170-8. doi:10.5115/acb.22.045
12. Stepanovsky M, Buk Z, Pilmann Koterova A, et al. Application of machine-learning methods in age-at-death estimation from 3D surface scans of the adult acetabulum. Forensic Sci Int. 2024; 365: 112272. doi:10.1016/j.forsciint.2024.112272
13. Imaizumi K, Usui S, Nagata T, et al. Sex-estimation method for three-dimensional shapes of the skull

and skull parts using machine learning. Forensic Sci Int. 2025; 373: 112532. doi:10.1016/j.forsciint.2025.112532
14. Faragalli A, Ferrante L, Angelakopoulos N, et al. Do machine learning methods solve the main pitfall of linear regression in dental age estimation? Forensic Sci Int. 2025; 367: 112353. doi:10.1016/j.forsciint.2024.112353
15. Parlak ME, Ozkul BB, Oruc M, et al. Sex and stature estimation from anthropometric measurements of the foot: linear analyses and neural network approach on a Turkish sample. Egypt J Forensic Sci. 2024; 14(1): 18. doi:10.1186/s41935-024-00391-4
16. Lye R, Min H, Dowling J, et al. Deep learning versus human assessors: forensic sex estimation from three-dimensional computed tomography scans. Sci Rep. 2024; 14(1): 30136. doi:10.1038/s41598-024-81718-y
17. Cao Y, Ma Y, Yang X, et al. Use of deep learning in forensic sex estimation of virtual pelvic models from the Han population. Forensic Sci Res. 2022; 7(3): 540-9. doi:10.1080/20961790.2021.2024369
18. Avent PR, Hughes CE, Garvin HM. Applying posterior probability informed thresholds to traditional cranial trait sex estimation methods. J Forensic Sci. 2022; 67(2): 440-9. doi:10.1111/1556-4029.14947
19. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007; 370(9596): 1453-7. doi:10.1016/S0140-6736(07)61602-X
20. Keereewan W, Monum T, Prasitwattanaseree S, et al. Stature estimation using the sacrum in a Thai population. Anat Cell Biol. 2023; 56(2): 259-67. doi:10.5115/acb.22.140
21. Martin R, Saller K. Lehrbuch der Anthropologie in systematischer Darstellung. 3rd ed. Stuttgart: Gustav Fischer Verlag; 1957.
22. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. Br J Nutr. 1999; 82(3): 165-77. doi:10.1017/S0007114599001348
23. Voulgari M, Anastopoulou I, Karakostis FA, et al. Testing the discrepancy between physical and virtual linear measurements. Anthropol Anz. 2024; 81(2): 153-9. doi:10.1127/anthranz/2023/1739
24. Ekizoglu O, Bozdag M, Er A, et al. Stature estimation in contemporary Turkish population: a CT study of lower limb long bones. Arch Leg Med. 2026; 17(2): 200662. doi:10.1016/j.aolm.2026.200662
25. Diac MM, Iov T, Damian SI, et al. Estimation of stature from tibia length for Romanian adult population. Appl Sci. 2021; 11(24): 11962. doi:10.3390/app112411962
26. Bidmos MA, Loubser N, Brits D. Regression equations for stature reconstruction from the lower limb bones of contemporary White South Africans. Aust J Forensic Sci. 2024; 56(4): 392-404. doi:10.1080/00450618.2023.2187084
27. Monum T, Opaburanakul T, Samai W, et al. Stature estimation using scapula measurements by postmortem computed tomography in the southern Thai population. Egypt J Forensic Sci. 2024; 14(1): 39. doi:10.1186/s41935-024-00413-1
28. Opaburanakul T, Monum T, Kosawiwat K, et al. Performance of the machine learning algorithm in stature estimation using scapula measurements from post-mortem computed tomography in a Thai male population. J Health Sci Med Res. 2025; 43(3): e20241118. doi:10.31584/jhsmr.20241118
29. Mittino G, Langstaff H, Garcia-Donas JG. Sex and stature estimation on the tibia: a virtual pilot study on a contemporary Hispanic population. J R Anthropol Inst. 2024; 30(3): 743-61. doi:10.1111/1467-9655.14103
30. Malawan P, Mahakkanukrauh P. Sex estimation from the tibia in Thais: validation of tibial tuberosity as a novel osteometric indicator. Egypt J Forensic Sci. 2026; 16(1): 65. doi:10.1186/s41935-026-00565-2
31. Demet Mutlu G, Asirdizer M, Kartal E, et al. Sex estimation from the hyoid bone measurements in an adult Eastern Turkish population using 3D CT images, discriminant function analysis, support vector machines, and artificial neural networks. Leg Med (Tokyo). 2024; 67: 102383. doi:10.1016/j.legalmed.2023.102383
32. Putri AS, Soedarsono N, Nehemia B, et al. Age estimation of individuals aged 5-23 years based on dental development of the Indonesian population. Forensic Sci Res. 2022; 7(2): 115-23. doi:10.1080/20961790.2021.1886648
33. Basjuri MK, Weinheimer AZ, Reza Adnan MRA, et al. Integrating nasal anthropometry with multivariate analyses to examine forensic population differentiation between Malaysian and Indonesian adults. Forensic Sci. 2026; 6(1): 14. doi:10.3390/forensicsci6010014
34. Barreiro MB, Fernandes J, Morgado M, et al. The tall order of stature estimation in burnt skeletal remains: a performance test in the Portuguese population. Int J Legal Med. 2025; 139(6): 2981-9. doi:10.1007/s00414-025-03567-2
35. Phetnui P, Poodendaen C, Choompoo N, et al. Independent validation of population-specific equations for sex and stature estimation from the humerus in Northeastern Thailand. Forensic Sci. 2026; 6(1): 1. doi:10.3390/forensicsci6010001